

AD \_\_\_\_\_

Award Number: DAMD17-00-1-0448

TITLE: Cox Model for Interval Censored Data in Breast Cancer  
Follow-Up Studies

PRINCIPAL INVESTIGATOR: George Y. C. Wong, Ph.D.

CONTRACTING ORGANIZATION: Strang Cancer Prevention Center  
New York, New York 10021-4601

REPORT DATE: July 2002

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command  
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;  
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

# REPORT DOCUMENTATION PAGE

Form Approved  
OMB No. 074-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE July 2002	3. REPORT TYPE AND DATES COVERED Annual (1 Jul 01 - 30 Jun 02)
4. TITLE AND SUBTITLE Cox Model for Interval Censored Data in Breast Cancer Follow-Up Studies		5. FUNDING NUMBERS DAMD17-00-1-0448
6. AUTHOR(S) George Y. C. Wong, Ph.D.		
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Strang Cancer Prevention Center New York, New York 10021-4601  E-MAIL: gwong@strang.org		8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Medical Research and Materiel Command Fort Detrick, Maryland 21702-5012		10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

20021230 209

12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for Public Release; Distribution Unlimited	12b. DISTRIBUTION CODE
---	------------------------

**13. ABSTRACT (Maximum 200 Words)**

The overall objective of this research proposal is semi-parametric inference of the Cox regression model for a survival function  $\Pr(X > x|Z = z) = S(x|z) = [S_o(x)]^{e^{z\beta}}$ , where  $X$  is subject to interval censoring,  $Z$  represents the covariates,  $S_o$  is a baseline survival function, and  $\beta$  represents the regression coefficients. One objective of our research is to develop asymptotic inference of the generalized maximum likelihood estimator (GMLE) of the regression coefficients  $\beta$  and  $S(\cdot|z)$ . A critical limitation with the GMLE approach under interval censoring is that it is computationally feasible only for a small data set. Thus the focus of another aspect of our research is the investigation of a simple alternative to the GMLE obtained by a two-step estimation procedure involving data grouping. In the second year of our research, we have established consistency results for the GMLE and the two-step estimators (TSE) of  $\beta$  and  $S(s|z)$ . The results will be useful to breast cancer researchers pursuing chemoprevention intervention trials involving surrogate endpoint biomarkers, and genetic epidemiologists conducting studies on familial aggregation of breast cancer and related cancers.

14. SUBJECT TERMS breast cancer, interval censored data, Cox regression model, generalized maximum likelihood, two-step estimation, consistency, asymptotic normality and efficiency			15. NUMBER OF PAGES 10
16. PRICE CODE			
17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT Unlimited

## FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

\_\_\_ Where copyrighted material is quoted, permission has been obtained to use such material.

\_\_\_ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

\_\_\_ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

N/A In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and use of Laboratory Animals of the Institute of Laboratory Resources, national Research Council (NIH Publication No. 86-23, Revised 1985).

N/A For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

N/A In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

N/A In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

N/A In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

*George Wong*  
*7/29/02*  
*Principal Investigator*

## A. TABLE OF CONTENTS

---

Front Cover	1
Report Documentation Page	2
Foreword	3
A. Table of contents	4
B. Introduction	5 – 6
C. Body	6 – 9
D. Key research accomplishments	9
E. Reportable outcomes	9
F. Conclusions	9
G. References	10

## B. INTRODUCTION

Interval-censored (IC) data are encountered in three areas of breast cancer research. The most common application is in clinical relapse follow-up studies in which the study endpoint is disease-free survival. When a patient relapses, it is usually known that the relapse takes place between two follow-up visits, and the exact time to relapse is unknown. In statistics, we say relapse time is interval censored. Interval censoring is also encountered in breast cancer registry studies in which information on family history of cancer is updated periodically. The Strang Breast Surveillance Program for women at increased risk for breast cancer, for instance, has enlisted over 800 women with complete pedigree information which is verified and updated continuously. Family history data such as age at diagnosis of a specific cancer, or a benign but risk-conferring condition, are obtained from each registrant at each update. Time to a cancer event, and definitely time to first detection of a benign condition, are at best known to fall in the time interval between the last update and age at diagnosis. A third but increasingly important area of application of interval censoring is in breast cancer chemoprevention experiments or prevention trials, which involve the observation of one or more surrogate endpoint biomarkers (SEB) over time. The scientific question of interest here is the estimation of time for the SEB to reach a target value, and time from cessation of intake of a chemopreventive agent to the loss of its protective effect. Unfortunately, the exact values of both these time variables are known only to lie in between two successive assay inspection times. In a breast cancer follow-up study, we will often encounter covariates (for instance, tumor size and nodal status in a relapse study, and baseline SEB value in a chemoprevention trial).

Let  $X$  denote a time-to-event variable with distribution  $F(x) = Pr(X \leq x)$ , or equivalently, survival function  $S(x) = 1 - F(x)$ . In interval censoring,  $X$  is not observed and is known only to lie in an observable interval  $(L, R)$ . In our previous DOD funded grant, we have made fundamental contributions to both the theory of the generalized maximum likelihood (GML) estimation of  $S$ , and the computation in connection with the inference of GML estimator (GMLE)  $\hat{S}$  of  $S$ . These contributions are restricted to the case of univariate interval-censored data without covariates.

The Cox proportional hazards model [1] specifies that covariates have a proportional effect on the hazard function of  $X$ . This model provides powerful means for fitting failure time observations to a distribution free model and for estimating the risk for failure associated with a vector of covariates. It is extensively used for right-censored data. Finkelstein [2] applied the Cox model to analysis of interval-censored data. However, she did not estab-

lish asymptotic properties of the GMLE of the parameters in the model and the approach is limited to small sample sizes due to the computational difficulty .

Our interest in IC data with covariates is driven by needs arising from two related areas of breast cancer research at Strang. First, our investigators in the Strang Cancer Genetics Program want to study various patterns of familial aggregation of breast, ovarian and other forms of cancer using family history data from the Strang Breast Surveillance Program. Studies of familial early onset of breast cancer, breast-ovarian and breast-prostate associations will lead to IC data with covariates; therefore, a proper statistical procedure together with a feasible software to deal with such data are very much needed. Second, we conducted a one-year chemoprevention trial of indole-3-carbinol (I3C) for breast cancer prevention. In this prevention trial we monitored the levels of two SEB's, a urinary estrogen metabolite ratio and a blood counterpart, both of which are subject to interval censoring. An earlier dose-ranging study of I3C conducted by Wong *et al* [3] has been published.

The overall aim of this research proposal is to develop statistical inference for interval-censored data with covariates that are encountered in breast cancer chemoprevention trials employing surrogate endpoint biomarkers, and in breast cancer registry follow-up studies of familial aggregation of breast and other forms of cancer. Asymptotic generalized maximum likelihood theory under the Cox regression model will be investigated and computer software package for maximum likelihood inference will be implemented.

## C. BODY

### C.1. Model Formulation and Likelihood Equations.

Let  $Y_{K,1} < Y_{K,2} < \dots < Y_{K,K}$  denote the follow-up times for a patient who has made  $K$  follow-up visits, in a longitudinal follow-up study. Since the number of visits for each patient may vary,  $K$  is a random positive integer. For convenience, define  $Y_{K,0} = 0$  and  $Y_{K,K+1} = \infty$ . The time-to-event variable of interest,  $X$ , is not directly observed; instead, it is known to lie in between two successive censoring time points  $(Y_{K,j}, Y_{K,j+1})$ , where  $j = 0, \dots, K$ . Note that  $X$  is left censored if  $j = 0$ , strictly interval censored if  $0 < j < K$ , and right censored if  $X > Y_{K,K}$ . The observable interval-censored data corresponding to  $X$  is given by

$$(L, R) = (Y_{K,i}, Y_{K,i+1}) \text{ if } Y_{K,i} < X \leq Y_{K,i+1}, i = 0, 1, \dots, K. \quad (1)$$

In addition to  $(L, R)$ , we also observe a  $p \times 1$  covariate vector  $Z$ . We assume that  $K$  and the  $Y_{k,j}$ 's are independent of  $(X, Z)$ .

The Cox regression model for the survival function at  $X = x$  given  $Z = z$  is represented by

$$S(x|z) = [S_o(x)]^{e^{z\beta}}, \quad (2)$$

where  $z\beta$  is the dot product of  $Z$  and  $\beta$ ,  $S_o(x)$  is a baseline survival function and  $\beta$  is a  $p$ -dimensional regression coefficient vector.

Let  $I_i = (L_i, R_i, z_i)$ ,  $i = 1, \dots, n$ , be a random sample of size  $n$  interval-censored observations with covariates. In terms of the original observed intervals, the likelihood function of  $S$  and  $b$  is given by

$$\mathbb{L} = \prod_{i=1}^n ((S(L_i))^{e^{bz_i}} - (S(R_i))^{e^{bz_i}}), \quad (3)$$

where  $S$  is a survival function, and  $b$  is a  $p \times 1$  dimensional vector. The GMLE of  $(S_o, \beta)$  is a value  $(S, b)$  that maximizes (3) over all survival functions  $S$  and all  $b \in \mathcal{R}^p$ .

Since  $S_o$  places all probability mass on the innermost intervals of the  $I_i$ 's (see Peto [4] or Turnbull [5]), it is often computationally simpler to express  $L$  in terms of innermost intervals.

We say that an interval  $A$  is an innermost interval of the  $I_i$ 's if  $A$  is a nonempty finite intersection of one or more of the  $I_i$ 's such that either  $I_i \cap A = \emptyset$  or  $I_i \cap A = A$  for each  $i$ . Suppose there are a total of  $m$  distinct innermost intervals  $A_i = (\xi_i, \eta_i]$ , where  $\eta_i \leq \xi_{i+1}$  and  $m \leq n$ . Then the likelihood function (3) is equivalently given by

$$\mathbb{L} = \prod_{i=1}^n [(\sum_{k>l_i} s_k)^{e^{z_i b}} - (\sum_{k>r_i} s_k)^{e^{z_i b}}], \quad (4)$$

where  $l_i = \sup\{j : \eta_j \leq L_i\}$ ,  $r_i = \sup\{j : \eta_j \leq R_i\}$  and  $s = (s_1, \dots, s_m)$  denote the vector of the probability weights. The log likelihood of  $(s, b)$  is

$$\mathcal{L}(s, b) = \sum_{i=1}^n \ln[(\sum_{k>l_i} s_k)^{e^{z_i b}} - (\sum_{k>r_i} s_k)^{e^{z_i b}}]. \quad (5)$$

Note that  $(\sum_{k>r_i} s_k)^{e^{z_i b}} = 1$  if  $r_i = 0$  and  $(\sum_{k>l_i} s_k)^{e^{z_i b}} = 0$  if  $l_i = m$ .

## C.2. Generalized maximum likelihood estimation.

A GMLE of  $(s, \beta)$  is a value of  $(s, b)$  that maximizes the likelihood function (5). We could follow the Newton-Raphson (NR) algorithm taken by Finkelstein [2]. However, this

would involve the inverse of a matrix of order  $(m + p - 1) \times (m + p - 1)$ . Since  $m$  can be potentially large when  $n$  is large, the NR algorithm is not feasible for a large data set.

We advocate a computationally simple approach by first grouping the original data  $(L_i, R_i)$  and then applying a two-step iterative scheme to obtain the two-step estimators (TSE) of  $S_o$  and  $\beta$  based on the innermost intervals corresponding to the grouped intervals.

In the first year of our research, we have successfully implemented a computer software to calculate the TSE's of  $S_o$  and  $\beta$ . A manuscript focusing on the two-step computation scheme is near completion and will be submitted for publication by the end of summer.

In our second year of research, we have applied our two-step estimation procedure to the Cox regression analysis of a long-term prognostic follow-up study involving 375 women with unilateral T1-2N0, T1-2N1 and T3-4 breast cancer. All the patients were treated at Memorial Sloan Kettering Cancer Center and the follow-up are being conducted at Strang Cancer Prevention Center. The main objective of the study is to assess the prognostic significance of bone marrow micrometastasis (BMM) in predicting relapse. Standard clinical variables including nodal status and tumor diameter were included in the Cox model. Although we have not yet established asymptotic normality to validate the P values that were reported for the study, our two-step Cox regression analysis gave strong indication that BMM was not as predictive of relapse as previously expected (Osborne and Wong [6]). We shall return to the BMM analysis when we fully establish the asymptotic normality of the GMLE of the Cox regression parameters. In our second year of research, therefore, we have moved ahead of our statement of work by making a start for Task 8. Since the BMM relapse follow-up study provides a complete and final data set that optimally satisfies our need of an empirical example to illustrate our asymptotic GML procedure for Cox regression, we instead to focus on this data set instead of the example mentioned in Task 8.

Also, in the second year of our research, we have established consistency of the GMLE of  $\beta$  and  $S_o$  (and hence  $S(\cdot|z)$ ) under the following assumptions:

AS1:  $S_o$  is arbitrary and each of the censoring variables,  $Y_1, \dots, Y_k$  takes on finitely many values.

AS2:  $S_o$  is arbitrary and each of the censoring variables,  $Y_1, \dots, Y_k$  is continuous and some regularity conditions are imposed on either  $S_o$  or the joint distribution function  $G$  of  $K, Y_1, \dots, Y_K$ .

Specifically, under AS1 and AS2

$$Pr\{\lim_{n \rightarrow \infty} \hat{\beta} = \beta\} = 1, \quad (6)$$

and

$$Pr\{\lim_{n \rightarrow \infty} \sup_{t \in H} |\hat{S}_o(t) - S_o(t)| = 0\} = 1, \quad (7)$$

where  $H$  denotes the support set of  $Y_1, \dots, Y_K$ . Note that  $\hat{S}_o(t)$  is guaranteed to be consistent for  $t \in H$ , and not elsewhere. However, the set  $H$  is not necessarily a time interval (for instance,  $H$  may be a collection of discrete points). In order for the consistency results to be more useful, we have established that if  $S_o$  is continuous, and the support of  $Y_1, \dots, Y_K$  is dense in  $[0, T]$  for some  $T > 0$ , then  $\hat{S}_o(t)$  is consistent for all  $t \in [0, T]$ . The practical implication of the denseness requirement is that pointwise consistency of  $\hat{S}_o(t)$  would hold only if all the subjects in a follow-up study must be followed at very frequent close intervals.

We have also established similar consistency results for the TSE, with an added assumption that the maximal length of the partition interval tends to 0 as  $n$  tends to  $\infty$ . The consistency results are being summarized in a manuscript under preparation.

#### D. KEY RESEARCH ACCOMPLISHMENTS

- We have completed Task 3 and Task 4 pertaining to consistency of GMLE of  $S_o(t)$  and  $\beta$ , and also consistency of TSE for the same parameters.
- We have made a start of Task 8 by performing a preliminary Cox regression analysis on a long-term prognostic relapse follow-up study involving 375 breast cancer patients.

#### E. REPORTABLE OUTCOMES

- An abstract presented at 2002 ASCO Meeting and published in the proceedings [3].
- A computer program to calculate the GMLE of the baseline survival function  $S_o$  and the Cox regression coefficients  $\beta$ .
- A computer program to calculate the TSE of  $S_o$  and  $\beta$ .

Both of these computer programs have been made available for the public via the Internet site <http://www.math.binghamton.edu/qyu/index.html>.

#### F. CONCLUSIONS

In the second year of our DOD grant, we have successfully completed our research objectives stated in Tasks 3 and 4. We have demonstrated that the GMLE and the TSE of the regression coefficients is consistent.

The results which we have established will be useful to breast cancer researchers pursuing chemoprevention intervention trials involving surrogate endpoints biomarkers, and genetic epidemiologists conducting studies on familial aggregation of breast cancer and related cancers.

## G. REFERENCES

- [1] Cox, D.R. (1972). Regression models and life tables. *J. Roy. Statist. Soc. B*, 34 187-220.
- [2] Finkelstein, D.M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics*, 42 845-854.
- [3] Wong, GY, Bradlow, HL, Sepkovic, D, Mehl, S, Mailman, J, and Osborne, MP (1997). A dose-ranging study of indole-3-carbinol for breast cancer prevention. *Journal of Cellular Biochemistry Supplement* 28/29 111-116.
- [4] Peto, R. (1973). Experimental survival curves for interval-censored data. *Appl. Statist.* 22, 86-91.
- [5] Turnbull, B. W. (1976). The empirical distribution function with arbitrary grouped, censored and truncated data. *J. Roy. Statist. Soc. Ser. B*, 38, 290-295.
- [6] Osborne, MP and Wong GYC. (2002). Breast cancer bone marrow micrometastases: a long-term prognostic study of systemic tumor cell burden on relapse. *Proceedings of American Society of Clinical Oncology*, 21, #228.